Nonparametric Bayesian Identification of Jump Systems with Sparse Dependencies ⋆

Emily B. Fox * **Erik B. Sudderth** ** **Michael I. Jordan** **
Alan S. Willsky *

* *MIT, Cambridge, MA 02139 USA (e-mail: {ebfox,willsky}@mit.edu)*
** *University of California, Berkeley 94720 USA (e-mail: {sudderth,jordan}@eecs.berkeley.edu)*

Abstract: Many nonlinear dynamical phenomena can be effectively modeled by a system that switches among a set of conditionally linear dynamical modes. We consider two such Markov jump linear systems: the switching linear dynamical system (SLDS) and the switching vector autoregressive (S-VAR) process. In this paper, we present a nonparametric Bayesian approach to identifying an unknown number of persistent, smooth dynamical modes by utilizing a hierarchical Dirichlet process prior. We additionally employ automatic relevance determination to infer a sparse set of dynamic dependencies. The utility and flexibility of our models are demonstrated on synthetic data and a set of honey bee dances.

Keywords: Jump processes; Non-parametric identification; Stochastic realization; Machine learning; Dynamic systems; Markov models; State-space models; Autoregressive processes.

1. INTRODUCTION

The *switching linear dynamical system* (SLDS) has been used to describe a multitude of nonlinear dynamical phenomena including human motion (Pavlović et al. [2000]), financial time series (Carvalho and Lopes [2006]), maneuvering targets (Rong Li and Jilkov [2005]), and honey bee dances (Oh et al. [2008]). The different dynamical modes account for structural changes the phenomena exhibit: a coasting ballistic missile makes an evasive maneuver; a country experiences a recession; a honey bee changes from a *waggle* to a *turn right* dance. Some of these changes appear frequently, while others are rarely observed, and there is always the possibility of a previously unseen dynamical behavior. These considerations motivate our nonparametric Bayesian approach, specifically employing hierarchical extensions of the Dirichlet process (DP), leading to flexible and computationally efficient learning of SLDS. While the DP aims at inferring a small set of representative dynamical modes, we also present a method of inducing sparsity in the dependency structure among variables, providing insight into the variable-order structure of the LDS.

One can view the SLDS, and a simpler *switching vector autoregressive* (S-VAR) process, as an extension of hidden Markov models (HMMs) in which each HMM state, or *mode*, is associated with a linear dynamical process. While the HMM makes a strong Markovian assumption that observations are conditionally independent given the mode, the SLDS and S-VAR are able to capture more complex temporal dependencies often present in real data. Most existing methods for learning SLDS and S-VAR rely on either fixing the number of HMM modes, such as in Oh et al. [2008], or considering a change-point detection formulation where each inferred change is to a new, previously unseen dynamical mode, such as in Xuan and Murphy [2007]. In this paper we show how one can remain agnostic about the number of dynamical modes while still allowing for returns to previously exhibited dynamical behaviors.

Hierarchical Dirichlet processes (HDP) can be used as a prior on the parameters of HMMs with unknown mode space cardinality (Beal et al. [2002], Teh et al. [2006]). In this paper we use a variant of the HDP-HMM—the *sticky HDP-HMM* of Fox et al. [2008a]—that provides improved control over the number of modes inferred; such control is crucial for the problems we examine. An extension of the sticky HDP-HMM formulation to learning SLDS and S-VAR with an unknown number of modes was presented in Fox et al. [2008b], however, relying on knowledge of the model order. In this paper, we explore a method for learning which components of the underlying state vector contribute to the dynamics of each mode by employing *automatic relevance determination* (ARD) (Beal [2003]). The resulting model allows for learning realizations of SLDS that switch between an unknown number of dynamical modes with possibly varying state dimensions, or S-VAR with varying autoregressive orders.

Paoletti et al. [2007] provides a survey of recent approaches to identification of switching dynamical models. For noiseless S-VAR, Vidal et al. [2003] presents an exact algebraic approach. However, the method relies on fixing the maximal mode space cardinality and autoregressive order, which is assumed shared between modes. Additionally, extensions to the noisy case rely on heuristics. Psaradakis and Spagnolo [2006] alternatively consider a penalized likelihood approach to identification of stochastic S-VAR. For SLDS, identification is significantly more challenging, and methods such as those in (Huang et al. [2004], Petreczky and Vidal [2007]) rely on simplifying assumptions such as deterministic dynamics or knowledge of the mode space. Extensions to standard SLDS would be quite challenging; our nonparametric framework further complicates these issues. The approach we present herein aims to address

⋆ This work was supported in part by MURIs funded through AFOSR Grant FA9550-06-1-0324 and ARO Grant W911NF-06-1-0076.

identification of mode space cardinality and model order of SLDS and S-VAR within a fully Bayesian framework.

2. BACKGROUND: DIRICHLET PROCESSES AND THE STICKY HDP-HMM

A Dirichlet process (DP), denoted by $\mathrm{DP}(\gamma H)$, is a distribution over random probability measures

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \qquad \theta_k \sim H \tag{1}$$

on a parameter space Θ, with H a base measure on Θ. The weights are sampled via a *stick-breaking construction*:

$$\beta_k = \beta'_k \prod_{\ell=1}^{k-1} (1 - \beta'_\ell) \qquad \beta'_k \sim \mathrm{Beta}(1, \gamma). \tag{2}$$

We denote this distribution by $\beta \sim \mathrm{GEM}(\gamma)$. The DP is commonly used as a prior on the parameters of a mixture model of unknown complexity, resulting in a *DP mixture model* (see Fig.1(a)). To generate observations, we choose $z_i \sim \beta$ and $y_i \sim F(\theta_{z_i})$.

The *hierarchical Dirichlet process* (HDP) (Teh et al. [2006]) extends the DP to cases in which groups of data are produced by related, yet distinct, generative processes. Taking a hierarchical Bayesian approach, the HDP places a global DP prior $\mathrm{DP}(\alpha G_0)$ on Θ, and then draws group specific distributions $G_j \sim \mathrm{DP}(\alpha G_0)$. Here, the base measure G_0 acts as an "average" distribution ($E[G_j|G_0] = G_0$) encoding the frequency of each shared, global parameter. This layering of DPs produces group specific distributions:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} \qquad \pi_j \sim \mathrm{DP}(\alpha\beta)\,. \tag{3}$$

To generate observation y_{ji}, the i^{th} observation in group j, we choose $z_{ji} \sim \pi_j$ and $y_{ji} \sim F(\theta_{z_{ji}})$. See Fig. 1(b).

It can be shown that the following finite, hierarchical mixture model converges in distribution to the HDP as $L \to \infty$ (Ishwaran and Zarepour [2002], Teh et al. [2006]):

$$\beta \sim \mathrm{Dir}(\gamma/L, \ldots, \gamma/L) \qquad \pi_j \sim \mathrm{Dir}(\alpha\beta_1, \ldots, \alpha\beta_L). \tag{4}$$

This *weak limit* approximation is used in Sec. 3.2.

The HDP can be used to develop an HMM with potentially infinite mode space (Teh et al. [2006]). For this *HDP-HMM*, each HDP group-specific distribution π_j is a mode-specific transition distribution. Here, however, observations are not pre-partitioned into groups. Let z_t denote the mode of the Markov chain at time t. The Markovian structure dictates $z_t \sim \pi_{z_{t-1}}$, implying that z_{t-1} indexes the group to which y_t is assigned (i.e., all observations with $z_{t-1} = j$ are assigned to group j). Due to the infinite mode space, there are infinitely many possible groups. The current HMM mode z_t then indexes the parameter θ_{z_t} used to generate observation y_t. See Fig. 1(c), ignoring the edges between the nodes representing the observations.

By sampling $\pi_j \sim \mathrm{DP}(\alpha\beta)$, the HDP prior encourages modes to have similar transition distributions ($E[\pi_{jk}|\beta] = \beta_k$). However, it does not differentiate self–transitions from moves between modes. When modeling dynamical processes with mode persistence, the flexible nature of the HDP-HMM prior allows for mode sequences with unrealistically fast dynamics to have large posterior probability. Recently, it has been shown (Fox et al. [2008a]) that one may mitigate this problem by instead considering a *sticky* HDP-HMM where π_j is distributed as follows:

$$\pi_j \sim \mathrm{DP}(\alpha\beta + \kappa\delta_j) \tag{5}$$

Here, $(\alpha\beta + \kappa\delta_j)$ indicates that an amount $\kappa > 0$ is added to the j^{th} component of $\alpha\beta$, thus increasing the expected probability of self-transition. When $\kappa = 0$ the original HDP-HMM is recovered. We place a vague prior on κ and learn the self-transition bias from the data.

3. THE HDP-SLDS AND HDP-AR-HMM MODELS

We recount the nonparametric Bayesian extensions of the SLDS and S-VAR developed in Fox et al. [2008b], first assuming the model order is known, and then presenting a new method for inferring this information from the data. These models are referred to as the *HDP-SLDS* and *HDP-AR-HMM*, respectively, with their generative processes summarized as follows (see Fig. 1(c)-(d)):

HDP-AR-HMM	HDP-SLDS
$z_t \sim \pi_{z_{t-1}}$	$z_t \sim \pi_{z_{t-1}}$
$\boldsymbol{y}_t = \sum_{i=1}^{r} A_i^{(z_t)} \boldsymbol{y}_{t-i} + \boldsymbol{e}_t(z_t)$	$\boldsymbol{x}_t = A^{(z_t)} \boldsymbol{x}_{t-1} + \boldsymbol{e}_t(z_t)$ $\boldsymbol{y}_t = C\boldsymbol{x}_t + \boldsymbol{v}_t$

(6)

Here, π_j is as defined in Sec. 2, r is a fixed autoregressive order, $\boldsymbol{e}_t(z_t) \sim \mathcal{N}(0, \Sigma^{(z_t)})$ and $\boldsymbol{v}_t \sim \mathcal{N}(0, R)$.

For these models, we place a prior on the dynamic parameters and compute their posterior from the data. For the HDP-SLDS we do, however, fix the measurement matrix, C, for reasons of identifiability. Let $\tilde{C} \in \mathbb{R}^{n \times d}$, $d \geq n$, be a full rank measurement matrix associated with a dynamical system defined by $\tilde{A}$. Then, without loss of generality, we may consider $C = [I_n\ 0]$ since there exists an invertible transformation T such that the pair $C = \tilde{C}T = [I_n\ 0]$ and $A = T^{-1}\tilde{A}T$ defines an equivalent system.[1] Our choice of the number of columns of zeros is, in essence, a choice of model order and one which we will address in Sec. 3.1.

3.1 Posterior Inference of Parameters $(A^{(k)}, \Sigma^{(k)}, R)$

In this section we focus on developing a prior to regularize the learning of different dynamical modes conditioned on a fixed mode assignment $z_{1:T}$. For the SLDS, we analyze the posterior distribution of the dynamic parameters given a fixed, known state sequence $\boldsymbol{x}_{1:T}$. Methods for learning the number of modes and resampling the sequences $\boldsymbol{x}_{1:T}$ and $z_{1:T}$ are discussed in Sec. 3.2.

We may write the dynamic equation generically as:

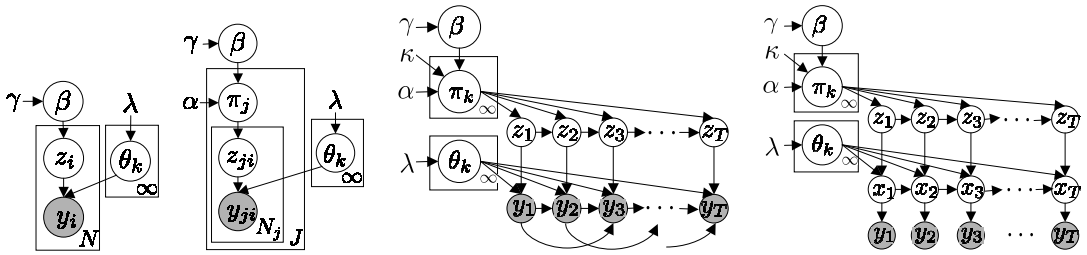
$$\boldsymbol{\psi_t} = \mathbf{A}^{(k)} \bar{\boldsymbol{\psi}}_{t-1} + \boldsymbol{e_t}. \tag{7}$$

For the S-VAR, we have $\mathbf{A}^{(k)} = [A_1^{(k)} \ldots A_r^{(k)}]$, $\boldsymbol{\psi_t} = \boldsymbol{y}_t$, and $\bar{\boldsymbol{\psi}}_t = [\boldsymbol{y}'_{t-1} \ldots \boldsymbol{y}'_{t-r}]'$. For the SLDS, $\mathbf{A}^{(k)} = A^{(k)}$, $\boldsymbol{\psi_t} = \boldsymbol{x}_t$, and $\bar{\boldsymbol{\psi}}_t = \boldsymbol{x}_{t-1}$.

Conditioned on the mode sequence, one may partition this dynamic sequence into K different linear regression problems, where $K = |\{z_1, \ldots, z_T\}|$. That is, for each mode k, we may form a matrix $\boldsymbol{\Psi}^{(k)}$ with N_k columns consisting of the $\boldsymbol{\psi}_t$ with $z_t = k$. Then,

$$\boldsymbol{\Psi}^{(k)} = \mathbf{A}^{(k)} \bar{\boldsymbol{\Psi}}^{(k)} + \mathbf{E}^{(k)}, \tag{8}$$

[1] For the SLDS of Eq. (6), T is identical for all modes as C is not mode-specific. This implies that for every realization $\mathcal{R}_1$ of the SLDS there exists an equivalent realization $\mathcal{R}_2$ with $C = [I_n\ 0]$.



(a) (b) (c) (d)

Fig. 1. For all graphs, $\beta \sim \text{GEM}(\gamma)$ and $\theta_k \sim H(\lambda)$. (a) DP mixture model in which $z_i \sim \beta$ and $y_i \sim f(y \mid \theta_{z_i})$. (b) HDP mixture model with $\pi_j \sim \text{DP}(\alpha, \beta)$, $z_{ji} \sim \pi_j$, and $y_{ji} \sim f(y \mid \theta_{z_{ji}})$. (c)-(d) Sticky HDP-HMM prior on switching VAR(2) and SLDS processes with the mode evolving as $z_{t+1} \sim \pi_{z_t}$ for $\pi_k \sim \text{DP}(\alpha + \kappa, (\alpha\beta + \kappa\delta_k)/(\alpha + \kappa))$. The dynamical processes are as in Eq. (6).

where $\bar{\boldsymbol{\Psi}}^{(k)}$ is a matrix of the associated $\bar{\boldsymbol{\psi}}_{t-1}$, and $\mathbf{E}^{(k)}$ the associated noise vectors.

Conjugate Prior The *matrix-normal inverse-Wishart* (MNIW) prior is conjugate for the parameter set $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$ (West and Harrison [1997]). A matrix $\boldsymbol{A} \in \mathbb{R}^{d \times m}$ has a matrix-normal distribution $\mathcal{MN}(\boldsymbol{A}; \boldsymbol{M}, \boldsymbol{V}, \boldsymbol{K})$ if

$$p(\boldsymbol{A}) = \frac{|\boldsymbol{K}|^{\frac{d}{2}}}{|2\pi \boldsymbol{V}|^{\frac{m}{2}}} e^{-\frac{1}{2} tr((\boldsymbol{A}-\boldsymbol{M})^T \boldsymbol{V}^{-1} (\boldsymbol{A}-\boldsymbol{M}) \boldsymbol{K})}, \quad (9)$$

where $\boldsymbol{M}$ is the mean matrix and $\boldsymbol{V}$ and $\boldsymbol{K}^{-1}$ are the covariances along the rows and columns, respectively.

Let $\mathbf{D}^{(k)} = \{\boldsymbol{\Psi}^{(k)}, \bar{\boldsymbol{\Psi}}^{(k)}\}$. The posterior distribution of the dynamic parameters for the k^{th} mode decomposes as

$$p(\mathbf{A}^{(k)}, \Sigma^{(k)} \mid \mathbf{D}^{(k)}) = p(\mathbf{A}^{(k)} \mid \Sigma^{(k)}, \mathbf{D}^{(k)}) p(\Sigma^{(k)} \mid \mathbf{D}^{(k)}). \quad (10)$$

The resulting posterior of $\mathbf{A}^{(k)}$ is derived to be

$$p(\mathbf{A}^{(k)} \mid \Sigma^{(k)}, \mathbf{D}^{(k)}) = \mathcal{MN}(\mathbf{A}^{(k)}; \mathbf{S}^{(k)}_{\psi\bar{\psi}} \mathbf{S}^{-(k)}_{\bar{\psi}\bar{\psi}}, \Sigma^{-(k)}, \mathbf{S}^{(k)}_{\bar{\psi}\bar{\psi}}), \quad (11)$$

with $\boldsymbol{B}^{-(k)}$ denoting $(\boldsymbol{B}^{(k)})^{-1}$ for a given matrix $\boldsymbol{B}$, and

$$\mathbf{S}^{(k)}_{\bar{\psi}\bar{\psi}} = \bar{\boldsymbol{\Psi}}^{(k)} \bar{\boldsymbol{\Psi}}^{(k)^T} + \boldsymbol{K} \qquad \mathbf{S}^{(k)}_{\psi\bar{\psi}} = \boldsymbol{\Psi}^{(k)} \bar{\boldsymbol{\Psi}}^{(k)^T} + \boldsymbol{M}\boldsymbol{K}$$
$$\mathbf{S}^{(k)}_{\psi\psi} = \boldsymbol{\Psi}^{(k)} \boldsymbol{\Psi}^{(k)^T} + \boldsymbol{M}\boldsymbol{K}\boldsymbol{M}^T.$$

Assuming an inverse-Wishart prior $\text{IW}(S_0, n_0)$ on $\Sigma^{(k)}$,

$$p(\Sigma^{(k)} \mid \mathbf{D}^{(k)}) = \text{IW}(\mathbf{S}^{(k)}_{\psi|\bar{\psi}} + S_0, N_k + n_0), \quad (12)$$

where $\mathbf{S}^{(k)}_{\psi|\bar{\psi}} = \mathbf{S}^{(k)}_{\psi\psi} - \mathbf{S}^{(k)}_{\psi\bar{\psi}} \mathbf{S}^{-(k)}_{\bar{\psi}\bar{\psi}} \mathbf{S}^{(k)^T}_{\psi\bar{\psi}}$.

Automatic Relevance Determination The MNIW prior leads to full $\mathbf{A}^{(k)}$ matrices, which (i) becomes problematic as the dimensionality of the underlying SLDS state, $\boldsymbol{x}_t$, grows in the presence of limited data; and (ii) does not provide a method for identifying irrelevant components of the state vector. To jointly address these issues, we alternatively consider *automatic relevance determination* (ARD) (Beal [2003], MacKay [1994]), which places independent, zero-mean, spherically symmetric Gaussian priors on the columns of the matrix $\mathbf{A}^{(k)}$:

$$p(\mathbf{A}^{(k)} | \alpha^{(k)}) = \prod_{j=1}^{d} \mathcal{N}(\boldsymbol{a}_j^{(k)}; 0, \alpha_j^{-(k)} I_d). \quad (13)$$

Each precision parameter $\alpha_j^{(k)}$ is given a $\text{Gamma}(a, b)$ prior. The zero-mean Gaussian prior penalizes non-zero columns of the dynamic matrix by an amount determined by the precision parameters. Iterative estimation of the precisions $\alpha_j^{(k)}$ and the dynamic matrix $\mathbf{A}^{(k)}$ leads to $\alpha_j^{(k)}$ becoming large, and thus $\boldsymbol{a}_j^{(k)}$ tending to 0, for columns whose evidence in the data is insufficient for overcoming the penalty induced by the prior. This implies that the j^{th} state component does not contribute to the dynamics of the k^{th} mode. Looking at the k^{th} dynamical mode alone, having $\boldsymbol{a}_j^{(k)} = 0$ implies that the realization of *that mode* is not minimal since a stochastic realization is minimal if and only if the following Hankel matrix has full rank:

$$\mathcal{H} = \left[C;\ C\boldsymbol{A};\ \cdots;\ C\boldsymbol{A}^{d-1}\right] \left[G\ \ \boldsymbol{A}G\ \cdots\ \boldsymbol{A}^{d-1}G\right], \quad (14)$$

where $G = \boldsymbol{A}P_x C^T$ and P_x is the steady-state covariance satisfying (assuming $\boldsymbol{A}$ stable),

$$E[\boldsymbol{x}_t \boldsymbol{x}_t^T] \triangleq P_x = \boldsymbol{A}P_x\boldsymbol{A}^T + Q. \quad (15)$$

However, the overall SLDS realization may still be minimal. In addition, each mode need not have stable dynamics for the SLDS to be stable (Costa et al. [2005]).

In this paper we restrict ourselves to ARD modeling of dynamical phenomena that satisfy the following criterion.

Criterion 1. If for some realization $\mathcal{R}$ a mode k has $\boldsymbol{a}_j^{(k)} = 0$, then $\boldsymbol{c}_j = 0$, where $\boldsymbol{c}_j$ is the j^{th} column of C. Thus, the set of observed state vector components is a subset of those relevant to *all* modes. We assume that the states are ordered such that $C = [C_0 \quad 0]$ (i.e., the observed components are the first components of the state vector.)

For example, if we have a 3-mode SLDS realization $\mathcal{R}$ with

$$\mathbf{A}^{(1)} = \left[\boldsymbol{a}_1^{(1)}\ \boldsymbol{a}_2^{(1)}\ \boldsymbol{a}_3^{(1)}\ 0\ 0\right] \quad \mathbf{A}^{(2)} = \left[\boldsymbol{a}_1^{(2)}\ \boldsymbol{a}_2^{(2)}\ 0\ \boldsymbol{a}_4^{(2)}\ 0\right]$$
$$\mathbf{A}^{(3)} = \left[\boldsymbol{a}_1^{(3)}\ \boldsymbol{a}_2^{(3)}\ \boldsymbol{a}_3^{(3)}\ 0\ \boldsymbol{a}_5^{(3)}\right], \quad (16)$$

then $C = [\boldsymbol{c}_1\ \boldsymbol{c}_2\ 0\ 0\ 0]$.

This criterion is sufficient, though not necessary, for maintaining the sparsity within each $\mathbf{A}^{(k)}$ while still fixing $C = [I_n\ 0]$. That is, given there exists a realization $\mathcal{R}_1$ of our dynamical phenomena that satisfies Criterion 1, the transformation T to an equivalent realization $\mathcal{R}_2$ with $C = [I_n\ 0]$ will maintain the sparsity structure seen in $\mathcal{R}_1$, which we aim to infer with the ARD prior. The above criterion is reasonable for many applications, as we often have observations of some components of the state vector that are essential to *all* modes, but *some* modes may have additional unobserved components that affect the dynamics. If there does not exist a realization $\mathcal{R}$ satisfying Criterion 1, we may instead consider a more general model where the measurement equation is mode-specific and we place a prior on $C^{(k)}$ instead of fixing this matrix. However, this model leads to identifiability issues that are considerably less pronounced in the above case.

The ARD prior may also be used to learn variable-order S-VAR processes. Instead of placing independent Gaussian priors on each column of $\mathbf{A}^{(k)}$, as we did in Eq. 13 for the SLDS, we decompose the prior over the *lag blocks* $A_i^{(k)}$:

$$p(\mathbf{A}^{(k)}|\boldsymbol{\alpha}^{(k)}) = \prod_{i=1}^{r} \mathcal{N}(\mathrm{vec}(A_i^{(k)}); 0, \alpha_i^{-(k)} I_{n^2}). \quad (17)$$

That is, each block $A_i^{(k)}$ shares the same precision $\alpha_i^{(k)}$ so entire lag components of that mode can be "turned off".

Our ARD prior on $\mathbf{A}^{(k)}$ is equivalent to a $\mathcal{N}(0, \Sigma_0^{(k)})$ prior on $\mathrm{vec}(\mathbf{A}^{(k)})$, where

$$\Sigma_0^{(k)} = \mathrm{diag}\left(\alpha_1^{(k)}, \ldots, \alpha_1^{(k)}, \ldots, \alpha_m^{(k)}, \ldots, \alpha_m^{(k)}\right)^{-1}. \quad (18)$$

Here, $m = d$ for the SLDS with d replicates of each $\alpha_i^{(k)}$, and $m = r$ for the S-VAR with n^2 replicates of $\alpha_i^{(k)}$. To examine the posterior distribution of $\mathbf{A}^{(k)}$, we note that we may rewrite the state equation as,

$$\begin{aligned}\boldsymbol{\psi}_{t+1} &= \left[\, \bar{\boldsymbol{\psi}}_{t,1} I_\ell \;\; \bar{\boldsymbol{\psi}}_{t,2} I_\ell \;\; \cdots \;\; \bar{\boldsymbol{\psi}}_{t,\ell * r} I_\ell \,\right] \mathrm{vec}(\mathbf{A}^{(k)}) + \boldsymbol{e}_{t+1}(k) \\ &\triangleq \tilde{\Psi}_t \mathrm{vec}(\mathbf{A}^{(k)}) + \boldsymbol{e}_{t+1}(k) \qquad \forall\, t | z_t = k, \end{aligned} \quad (19)$$

where $\ell = d$ for the SLDS and $\ell = n$ for the S-VAR, from which derive the posterior distribution as

$$p(\mathrm{vec}(\mathbf{A}^{(k)}) \mid \mathbf{D}^{(k)}, \Sigma^{(k)}, \boldsymbol{\alpha}^{(k)}) = \mathcal{N}^{-1}\Bigg(\sum_{t|z_t=k} \tilde{\Psi}_{t-1}^T \Sigma^{-(k)} \boldsymbol{\psi_t}, \Sigma_0^{-(k)} + \sum_{t|z_t=k} \tilde{\Psi}_{t-1}^T \Sigma^{-(k)} \tilde{\Psi}_{t-1} \Bigg). \quad (20)$$

Here, $\mathcal{N}^{-1}(\theta, \Lambda)$ represents a Gaussian $\mathcal{N}(\mu, \Sigma)$ with information parameters $\theta = \Sigma^{-1}\mu$ and $\Lambda = \Sigma^{-1}$. The posterior of each precision parameter $\alpha_\ell^{(k)}$ is given by

$$p(\alpha_\ell^{(k)} \mid \mathbf{D}^{(k)}, \mathbf{A}^{(k)}) = \mathrm{Gamma}\left(a + \frac{|\mathcal{S}_\ell|}{2}, b + \sum_{(i,j)\in\mathcal{S}_\ell} \frac{a_{ij}^{(k)^2}}{2} \right). \quad (21)$$

$\mathcal{S}_\ell$ are the indices for which $a_{ij}^{(k)}$ has prior precision $\alpha_\ell^{(k)}$.

We then place a separate inverse-Wishart prior $\mathrm{IW}(S_0, n_0)$ on $\Sigma^{(k)}$ and look at the posterior given $\mathbf{A}^{(k)}$:

$$p(\Sigma^{(k)} \mid \mathbf{D}^{(k)}, \mathbf{A}^{(k)}) = \mathrm{IW}(\mathbf{S}_{\psi|\bar{\psi}}^{(k)} + S_0, N_k + r_0), \quad (22)$$

where $\mathbf{S}_{\psi|\bar{\psi}}^{(k)} = \sum_{t|z_t=k} (\boldsymbol{\psi}_t - \mathbf{A}^{(k)}\bar{\boldsymbol{\psi}}_{t-1})(\boldsymbol{\psi}_t - \mathbf{A}^{(k)}\bar{\boldsymbol{\psi}}_{t-1})^T$.

Posterior of Measurement Noise For the HDP-SLDS, we additionally place an $\mathrm{IW}(R_0, r_0)$ prior on the measurement noise covariance R, which is shared between modes. The posterior distribution is given by

$$p(R \mid \boldsymbol{y}_{1:T}, \boldsymbol{x}_{1:T}) = \mathrm{IW}(S_R + R_0, T + r_0), \quad (23)$$

where $S_R = \sum_{t=1}^T (\boldsymbol{y}_t - C\boldsymbol{x}_t)(\boldsymbol{y}_t - C\boldsymbol{x}_t)^T$.

3.2 Gibbs Sampler

For inference in the HDP-AR-HMM, we use a Gibbs sampler that iterates between sampling the mode sequence, $z_{1:T}$, and the set of dynamic and sticky HDP-HMM parameters. See Fox et al. [2008a] for details of sampling $(\{\pi_k\}, \beta, \alpha, \kappa, \gamma)$ given $z_{1:T}$. The sampler for the HDP-SLDS is identical with the additional step of sampling the state sequence, $\boldsymbol{x}_{1:T}$, and conditioning on the state sequence when resampling dynamic parameters.

Sampling $(A^{(k)}, \Sigma^{(k)}, R)$ Conditioned on the mode sequence, $z_{1:T}$, and $\boldsymbol{\psi}_{1:T}$ (i.e., the observations, $\boldsymbol{y}_{1:T}$, or sampled state sequence, $\boldsymbol{x}_{1:T}$), we can sample the dynamic parameters $\boldsymbol{\theta} = \{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$ from the posterior densities of Sec. 3.1. For the ARD prior, we then sample $\boldsymbol{\alpha}^{(k)}$ given $\mathbf{A}^{(k)}$. For the HDP-SLDS, we additionally sample R.

Sampling $z_{1:T}$ As shown in Fox et al. [2008a], the mixing rate of the Gibbs sampler for the HDP-HMM can be dramatically improved by using a truncated approximation to the HDP, such as the weak limit approximation, and jointly sampling the mode sequence using a variant of the forward-backward algorithm. Specifically, we compute backward messages $m_{t+1,t}(z_t) \propto p(\boldsymbol{\psi}_{t+1:T} | z_t, \bar{\boldsymbol{\psi}}_t, \boldsymbol{\pi}, \boldsymbol{\theta})$ and then recursively sample each z_t conditioned on z_{t-1} from

$$\begin{aligned} p(z_t \mid z_{t-1}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) &\propto p(z_t \mid \pi_{z_{t-1}}) \\ &\quad p(\boldsymbol{\psi}_t \mid \bar{\boldsymbol{\psi}}_{t-1}, \mathbf{A}^{(z_t)}, \Sigma^{(z_t)}) m_{t+1,t}(z_t), \end{aligned} \quad (24)$$

Joint sampling of the mode sequence is especially important when the observations are directly correlated via a dynamical process since this correlation further slows the mixing rate of the sampler of Teh et al. [2006]. Note that the approximation of Eq. (4) retains the HDP's nonparametric nature by encouraging the use of fewer than L components while allowing the generation of new components, upper bounded by L, as new data are observed.

Sampling $\boldsymbol{x}_{1:T}$ *(HDP-SLDS only)* Conditioned on the mode sequence $z_{1:T}$ and the set of dynamic parameters $\boldsymbol{\theta}$, our dynamical process simplifies to a time-varying linear dynamical system. We can then block sample $\boldsymbol{x}_{1:T}$ by first running a backward filter to compute $m_{t+1,t}(\boldsymbol{x}_t) \propto p(\boldsymbol{y}_{t+1:T} | \boldsymbol{x}_t, z_{t+1}, \boldsymbol{\theta})$ and then recursively sampling each $\boldsymbol{x}_t$ conditioned on $\boldsymbol{x}_{t-1}$ from

$$\begin{aligned} p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) &\propto p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, A^{(z_t)}, \Sigma^{(z_t)}) \\ &\quad p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, R) m_{t+1,t}(\boldsymbol{x}_t). \end{aligned} \quad (25)$$

The messages are given in information form by $m_{t,t-1}(\boldsymbol{x}_{t-1}) \propto \mathcal{N}^{-1}(\boldsymbol{x}_{t-1}; \vartheta_{t,t-1}, \Lambda_{t,t-1})$, where the information parameters are recursively defined as

$$\begin{aligned} \vartheta_{t,t-1} &= A^{(z_t)^T} \Sigma^{-(z_t)} \tilde{\Lambda}_t (C^T R^{-1} \boldsymbol{y}_t + \vartheta_{t+1,t}) \\ \Lambda_{t,t-1} &= A^{(z_t)^T} \Sigma^{-(z_t)} A^{(z_t)} - A^{(z_t)^T} \Sigma^{-(z_t)} \tilde{\Lambda}_t \Sigma^{-(z_t)} A^{(z_t)}, \end{aligned} \quad (26)$$

where $\tilde{\Lambda}_t = (\Sigma^{-(z_t)} + C^T R^{-1} C + \Lambda_{t+1,t})^{-1}$.

4. RESULTS

MNIW prior We begin by analyzing the relative modeling power of the HDP-VAR(1)-HMM,[2] HDP-VAR(2)-HMM, and HDP-SLDS using the MNIW prior on three sets of test data (see Fig. 2.) We compare to a baseline sticky HDP-HMM using first difference observations, imitating a HDP-VAR(1)-HMM with $A^{(k)} = I$ for all k. The MNIW hyperparameters are set from statistics of the data. Our Hamming distance error metric is calculated by choosing the optimal mapping of indices maximizing overlap between the true and estimated mode sequences. For the first scenario, the data were generated from a 5-mode S-VAR(1) process. The three switching linear dynamical models provide comparable performance since both the HDP-VAR(2)-HMM and HDP-SLDS with $C = I$ contain

[2] Here we use the notation HDP-VAR(r)-HMM to refer to a HDP-AR-HMM with autoregressive order r and vector observations.

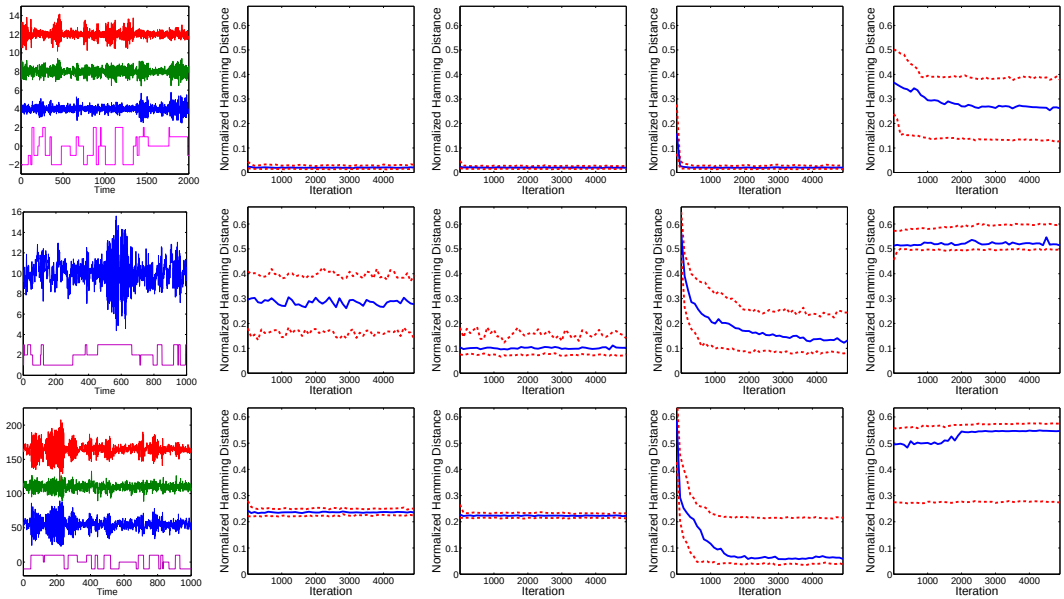


Fig. 2. (a) Observation sequence (blue, green, red) and associated mode sequence (magenta) for a 5-mode switching VAR(1) process (top), 3-mode switching AR(2) process (middle), and 3-mode SLDS (bottom). The associated 10th, 50th, and 90th Hamming distance quantiles over 100 trials are shown for the (b) HDP-VAR(1)-HMM, (c) HDP-VAR(2)-HMM, (d) HDP-SLDS with $C = I$ (top and bottom) and $C = [1\ \ 0]$ (middle), and (e) sticky HDP-HMM using first difference observations.

the class of HDP-VAR(1)-HMMs. In the second scenario, the data were generated from a 3-mode S-AR(2) process. The HDP-AR(2)-HMM has significantly better performance than the HDP-AR(1)-HMM while the performance of the HDP-SLDS with $C = [1\ 0]$ performs similarly, but has greater posterior variability because the HDP-AR(2)-HMM model family is smaller. Note that the HDP-SLDS sampler is slower to mix since the hidden, continuous state is also sampled. The data in the third scenario were generated from a 3-mode SLDS model with $C = I$. Here, we clearly see that neither the HDP-VAR(1)-HMM nor HDP-VAR(2)-HMM is equivalent to the HDP-SLDS. Note that all of the switching models yielded significant improvements relative to the baseline sticky HDP-HMM. Together, these results demonstrate both the differences between our models as well as the models' ability to learn switching processes with varying numbers of modes.

ARD prior We now compare the utility of the ARD prior to the MNIW prior using the HDP-SLDS model when the true underlying dynamical modes have sparse dependencies relative to the assumed model order. We generated data from a two-mode SLDS with 0.98 probability of self-transition and the following dynamical parameters:

$$\mathbf{A}^{(1)} = \begin{bmatrix} 0.8 & -0.2 & 0 \\ -0.2 & 0.8 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \mathbf{A}^{(2)} = \begin{bmatrix} -0.2 & 0 & 0.8 \\ 0.8 & 0 & -0.2 \\ 0 & 0 & 0 \end{bmatrix}, \quad (27)$$

with $C = [I_2\ 0]$, $\Sigma^{(1)} = \Sigma^{(2)} = I_3$, and $R = I_2$. The first dynamical process can be equivalently described by just the first and second state components, while the second process only relies on a single state component.[3] (The third component is equivalent to additional process noise.)

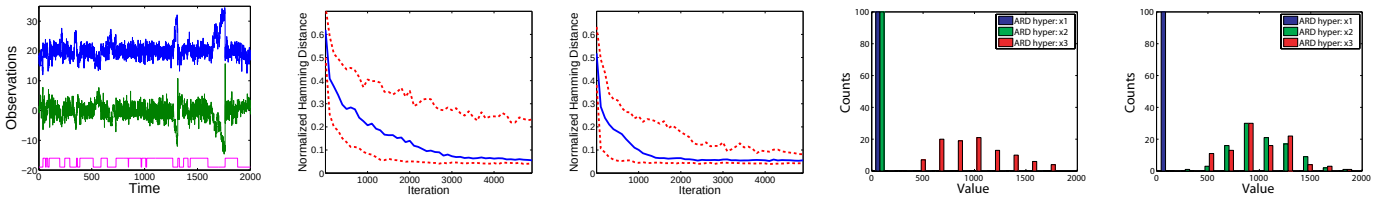
In Fig. 3, we see that the ARD provides superior mode-sequence estimates, as well as a mechanism for identifying non-dynamical state components. The histograms of Fig. 3(d)-(e) depict that we were able to correctly identify dynamical systems with $\boldsymbol{a}_3^{(1)} = 0$ and $\boldsymbol{a}_2^{(2)} = \boldsymbol{a}_3^{(2)} = 0$ since the precisions $\alpha_j^{(k)}$ corresponding to these columns are significantly larger than those for the other columns.

Dancing Honey Bees To analyze our ability to learn variable order S-VAR processes, we tested the HDP-VAR(2)-HMM model with the ARD prior on a set of three dancing honey bee sequences, aiming to segment the sequences into the *waggle*, *turn right*, and *turn left* dances displayed in Fig. 4(a). The data consist of measurements $\boldsymbol{y}_t = [\cos(\theta_t)\ \sin(\theta_t)\ x_t\ y_t]^T$, where (x_t, y_t) denotes the 2D coordinates of the bee's body and θ_t its head angle. Providing the data and ground truth labels, MATLAB's `lpc` implementation of Levinson's algorithm indicates that the turning dances are well approximated by an order 1 process, while the waggle dance relies on an order 2 model.[4] This same dataset was analyzed in Fox et al. [2008b] using the MNIW prior with the HDP-VAR(1)-HMM model, and compared against the 3-mode SLDS model of Oh et al. [2008] and the change-point detection formulation of Xuan and Murphy [2007].

The Hamming distance plots for the HDP-VAR(2)-HMM with the ARD prior, shown in Fig. 4(b), are indistinguishable from those using the HDP-VAR(1)-HMM with the MNIW prior. Thus, the information in the first lag component is sufficient for the segmentation problem. However, the ARD prior informs us of the variable-order nature of this switching dynamical process and would likely improve predictive performance of honey bee dances. From Fig. 4(d)-(f), we see that the learned orders for the three dances match what was indicated by running MATLAB's `lpc` function on ground-truth segmented data.

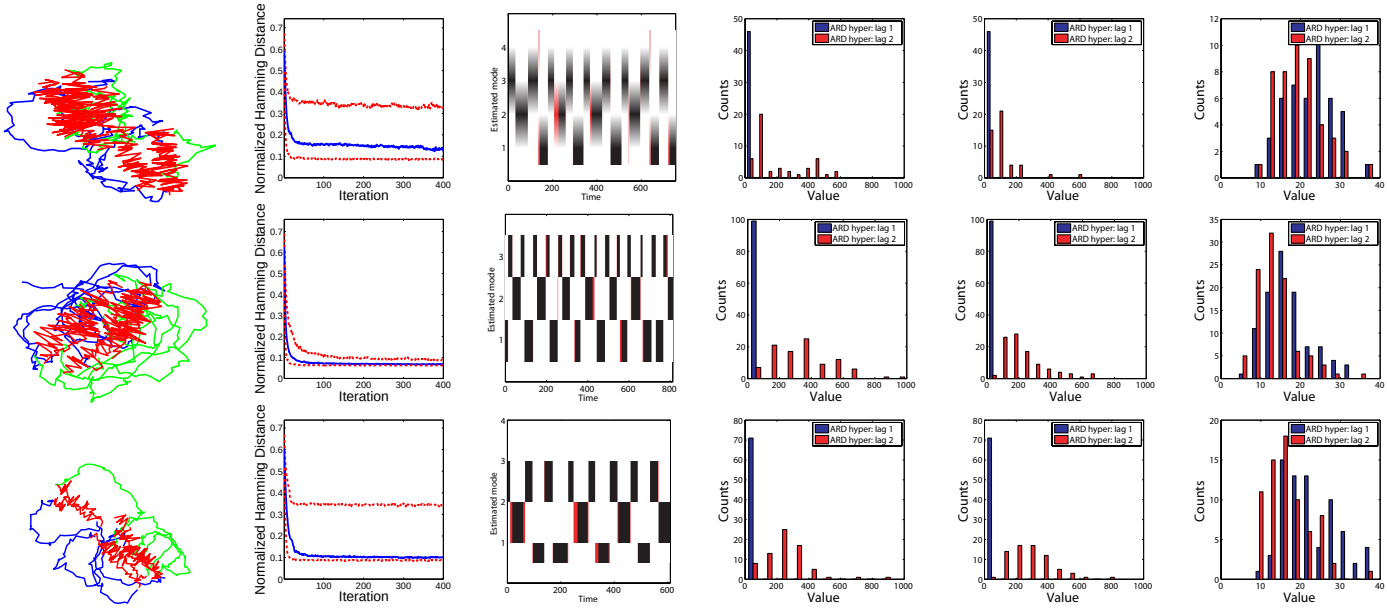
[3] Since the true model has $C = [I_2\ 0]$ rather than arbitrary $C = [C_0\ 0]$, we still hope to recover the $\boldsymbol{a}_2^{(2)} = \boldsymbol{a}_3^{(2)} = 0$ sparsity structure.

[4] `lpc` computes AR coefficients for scalar data, so we analyzed each component of the observation vector independently. The order was consistent across these components.



(a) (b) (c) (d) (e)

Fig. 3. (a) Observation sequence (green, blue) and mode sequence (magenta) of a 2-mode SLDS, where the first mode can be realized by the first two state components and the second mode solely by the first. The associated 10th, 50th, and 90th Hamming distance quantiles over 100 trials are shown for the (b) MNIW and (c) ARD prior. (d)-(e) Histograms of inferred ARD precisions associated with the first and second dynamical modes, respectively, at the 5000th Gibbs iteration. Larger values correspond to non-dynamical components.



(a) (b) (c) (d) - turn right (e) - turn left (f) - waggle

Fig. 4. (a) Honey bee trajectories colored by *waggle* (red), *turn right* (blue), and *turn left* (green) dances. (b) 10th, 50th, and 90th Hamming distance quantiles over 100 trials. (c) Estimated mode sequences, with errors in red, representing the median error at Gibbs iteration 200. (d)-(f) Histograms of inferred ARD precisions for the learned dance modes at Gibbs iteration 400 for the trials with Hamming distance below the median. Larger values correspond to unnecessary lag components. Note the horizontal axis scale in column (f).

5. DISCUSSION

In this paper, we have addressed the problem of learning switching linear dynamical models with an unknown number of modes for describing complex dynamical phenomena. We further examined methods of inferring sparse dependency structure in the dynamics of these modes, leading to flexible and scalable dynamical models. The utility of the developed HDP-SLDS and HDP-AR-HMM was demonstrated on simulated data, as well as sequences of honey bee dances in which we successfully inferred both segmentations of the data into *waggle*, *turn-right*, and *turn-left* dances as well as the VAR order of these dances.

REFERENCES

M. Beal. *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, University College London, London, UK, 2003.

M. Beal, Z. Ghahramani, and C. Rasmussen. The infinite hidden Markov model. In *NIPS*, 2002.

C. Carvalho and H. Lopes. Simulation-based sequential analysis of Markov switching stochastic volatility models. *Comp. Stat. & Data Anal.*, 2006.

O. L. V. Costa, M. V. Fragoso, and R. P. Marques. *Discrete-Time Markov Jump Linear Systems*. Springer, 2005.

E. Fox, E. Sudderth, M. Jordan, and A. Willsky. An HDP-HMM for systems with state persistence. In *ICML*, 2008a.

E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Nonparametric Bayesian learning of switching dynamical systems. In *NIPS*, 2008b.

K. Huang, A. Wagner, and Y. Ma. Identification of hybrid linear time-invariant systems via SES. In *CDC*, 2004.

H. Ishwaran and M. Zarepour. Exact and approximate sum–representations for the Dirichlet process. *Can. J. Stat.*, 30:269–283, 2002.

D.J.C. MacKay. *Bayesian methods for backprop networks*, chapter 6, pages 211–254. Models of Neural Networks, III. Springer, 1994.

S. Oh, J. Rehg, T. Balch, and F. Dellaert. Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. *Int. J. Comput. Vision*, 77(1–3):103–124, 2008.

S. Paoletti, A. Juloski, G. Ferrari-Trecate, and R. Vidal. Identification of hybrid systems: A tutorial. *Eur. J. Control*, 2–3:242–260, 2007.

V. Pavlović, J. Rehg, and J. MacCormick. Learning switching linear models of human motion. In *NIPS*, 2000.

M. Petreczky and R. Vidal. Realization theory of stochastic jump-Markov linear systems. In *CDC*, 2007.

Z. Psaradakis and N. Spagnolo. Joint determination of the state dimension and autoregressive order for models with Markov regime switching. *J. Time Series Anal.*, 27:753–766, 2006.

X. Rong Li and V. Jilkov. Survey of maneuvering target tracking. Part V: Multiple-model methods. *IEEE Trans. Aerosp. Electron. Syst.*, 41(4):1255–1321, 2005.

Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *J. Amer. Stat. Assoc.*, 101(476):1566–1581, 2006.

R. Vidal, S. Soatto, Y. Ma, and S. Sastry. An algebraic geometric approach to the identification of a class of linear hybrid systems. In *CDC*, 2003.

M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, 1997.

X. Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series. In *ICML*, 2007.